

Aus der US 5,317,646 geht ein Verfahren hervor, das einen Bediener beim Erzeugen von elektronischen Vorlagen für ein Formularerkennungssystem unterstützt. Hierzu wird ein nicht mit Daten versehenes Formular, ein sogenanntes Master-Formular, an einem Bildschirm dargestellt und der Benutzer kann die Datenfelder mit einer Zeigereinrichtung kennzeichnen. Die Koordinaten, die den entsprechenden Bereich begrenzen, werden automatisch erfasst, nachdem ein einziger Punkt innerhalb dieses Bereiches vom Bediener ausgewählt worden ist. Mit diesem Verfahren können schnell und einfach Vorlagen für die automatische Formularerkennung erstellt werden.

In Casey R.G. et al., „Intelligent Forms Processing“, IBM Systems Journal vol. 29 (1990) Nr. 3, Seiten 435 bis 450 wird ein Formularerkennungsverfahren beschrieben, bei dem ein eingescanntes Formular mittels Bildverarbeitungstechniken analysiert und mit anderen gespeicherten Vorlageformularen verglichen wird. Falls keine Übereinstimmung mit einem Vorlageformular gefunden wird, muss ein neues Vorlageformular durch Eingabe am Computer erzeugt werden. Beim Erzeugen einer Vorlage wird das gescannte Formular am Bildschirm dargestellt und mit einer Zeigereinrichtung werden die Begrenzungslinien der Eingabefelder markiert.

Aus der US 2002/141660 A1 geht ein zweistufiges Verfahren hervor, bei welchem zunächst Formularvorlagen eingegeben werden können und danach anhand der eingegebenen Formularvorlagen Dokumente automatisch ausgelesen werden können. Einzugebende Formularvorlagen werden gescannt und der Bediener zeigt mit einem Cursor auf Eingabefelder. Die Position und Größe der Eingabefelder wird gespeichert. Der Bediener kann auch den jedem Datenfeld zugeordneten Datentyp bestimmen. Beim automatischen Lesen von Formularen werden diese eingescannt und

anhand der in den gespeicherten Formularvorlagen enthaltenen Datenfeldern automatisch ausgelesen. Falls beim Auslesen ein Fehler auftritt, kann der Bediener über die Tastatur den Fehler korrigieren.

5

Die US 6,028,970 betrifft ein Verfahren und ein System zur automatischen Texterkennung (OCR). Das System umfasst ein Fehlerkorrigiermodul („error correction logic module“). Dieses Fehlerkorrigiermodul wird auf klar erkennbare Datenfehler angewandt, um diese zu korrigieren. Diese Korrekturen werden automatisch ausgeführt. Hierbei werden nicht nur Fehler einzelner Buchstaben erkannt, sondern auch Fehler im Kontext analysiert und entsprechend korrigiert. Ein nicht automatisch korrigierbarer Fehler kann dem Bediener mittels einer Fehler-

10

15

nachricht mitgeteilt werden. Der Bediener kann dann den mittels der Texterkennung erzeugten Text beurteilen und gegebenenfalls korrigieren.

20

## Ansprüche

1. Verfahren zum Erfassen von Daten aus maschinell lesbaren Dokumenten, wobei die Daten einer Datenbank zugeordnet werden, indem einzelne Daten möglichst automatisch dem Dokument extrahiert werden und in entsprechende Datenbankfelder eingetragen werden, und falls Daten für ein oder mehrere bestimmte Datenbankfelder einem Dokument nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, werden folgende Schritte ausgeführt:
  - Darstellen des Dokumentes an einem Bildschirm,
  - Anzeigen des Datenbankfeldes, für das die Daten nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, am Bildschirm,
  - Ausführen einer Vorschlags-Routine, mit welcher Stringabschnitte in der Nähe eines von einem Benutzer auf dem Bildschirm bewegbaren Zeiger ausgewählt, markiert und zur Extraktion vorgeschlagen werden.
2. Verfahren nach Anspruch 1,  
d a d u r c h g e k e n n z e i c h n e t,  
dass der Stringabschnitt nach Maßgabe von dem Datenbankfeld zugeordneten Konzept-Informationen ausgewählt, markiert und zur Extraktion vorgeschlagen wird.
3. Verfahren nach Anspruch 2,  
d a d u r c h g e k e n n z e i c h n e t,  
dass die Konzept-Informationen die Syntax und/oder die Semantik des Datenbankfeldes beschreiben, so dass von der Vorschlags-Routine ein zu markierender Stringabschnitt entsprechend der Syntax bzw. der Semantik des jeweiligen Datenbankfeldes ausgewählt und markiert wird.

4. Verfahren nach Anspruch 3,  
dadurch gekennzeichnet,  
dass die Informationen zur Syntax die Anzahl von Ziffern  
und/oder Buchstaben und/oder vorbestimmte Formate des zu  
5 lesenden Stringabschnitts beschreiben.
5. Verfahren nach Anspruch 3 oder 4,  
dadurch gekennzeichnet,  
dass die Informationen zur Semantik vorbestimmte Begriffe,  
10 bspw. mit einem Lexikon, beschreiben.
6. Verfahren nach einem der Ansprüche 1 bis 5,  
dadurch gekennzeichnet,  
dass ein Stringabschnitt ausgewählt wird, der zwischen  
15 zwei Begrenzungszeichen angeordnet ist.
7. Verfahren nach Anspruch 6,,  
dadurch gekennzeichnet,  
dass Begrenzungszeichen Leerzeichen und/oder Satzzeichen  
20 umfassen.
8. Verfahren nach einem der Ansprüche 1 bis 7,  
dadurch gekennzeichnet,  
dass der Text von Dokumenten in grafischer Darstellung  
25 zunächst mit einem OCR-Verfahren in codierten Text umgesetzt  
wird und die Vorschlags-Routine zusätzlich zum markierten  
Stringabschnitt in grafischer Darstellung den codierten Text  
dieses Stringabschnittes darstellt.
- 30 9. Verfahren nach einem der Ansprüche 1 bis 7,  
dadurch gekennzeichnet,  
dass zusätzlich zum markierten Stringabschnitt dieser  
Stringabschnitt nochmals in vergrößerter Darstellung am  
Bildschirm angezeigt wird.

35

10. Verfahren nach einem der Ansprüche 1 bis 9,  
dadurch gekennzeichnet,  
dass die Vorschlags-Routine nach dem Markieren eines  
Stringabschnittes eine Funktion aktiviert, mit welcher  
5 durch Betätigen einer oder mehrerer vorbestimmter Tasten  
der Inhalt des markierten Stringabschnittes in die Daten-  
bank übernommen wird.
11. Verfahren nach einem der Ansprüche 1 bis 10,  
dadurch gekennzeichnet,  
dass beim Ausführen der Vorschlags-Routine nach dem Bewe-  
gen des Zeigers ein vorbestimmtes Zeitintervall abgewart-  
tet wird, innerhalb dessen der Zeiger nicht bewegt werden  
darf, bevor ein Stringabschnitt ausgewählt wird.
12. Verfahren zum Erfassen von Daten aus maschinell lesbaren  
Dokumenten, wobei die Daten einer Datenbank zugeordnet  
werden, insbesondere nach einem der Ansprüche 1 - 11,  
dadurch gekennzeichnet,  
20 dass nach dem Einlesen von Daten aus einer ersten Tabel-  
lenzeile in korrespondierende Datenbankfelder automatisch  
die weiteren Tabelleneinträge durch einen Vergleich von  
unterhalb der ersten Tabellenzeile angeordneten Stringab-  
schnitten mit den Stringabschnitten der ersten Tabellen-  
zeile, die den eingelesenen Daten entsprechen, ermittelt  
25 werden und diese weiteren Tabelleneinträge automatisch  
extrahiert werden.
13. Verfahren nach Anspruch 12,  
30 dadurch gekennzeichnet,  
dass der Vergleich zwischen den Stringabschnitten mittels  
eines String Matching-Verfahrens erfolgt.
14. Verfahren nach Anspruch 12 oder 13,  
35 dadurch gekennzeichnet,  
dass die ermittelten Tabelleneinträge markiert werden.

15. Verfahren nach Anspruch 14,  
d a d u r c h g e k e n n z e i c h n e t,  
dass Funktionen zum Editieren der markierten Tabellenein-  
träge bereitgestellt werden.

5

16. System zum Erfassen von Daten aus maschinell lesbaren Do-  
kumenten umfassend einen Computer (12) mit einer Spei-  
chereinrichtung (13) und eine CPU (14), wobei in der  
Speichereinrichtung (13) ein Softwareprodukt zum Ausfüh-  
ren des Verfahrens nach einem der Ansprüche 1 - 15 ge-  
speichert ist.

10

17. System nach Anspruch 16,  
d a d u r c h g e k e n n z e i c h n e t,  
dass das System eine Eingabeeinrichtung (17) in Form ei-  
ner Maus (6) und/oder Tastatur (7) aufweist.

15

18. System nach Anspruch 16 oder 17,  
d a d u r c h g e k e n n z e i c h n e t,  
dass das System einen Scanner (16) zum optischen Abtasten  
von Dokumenten aufweist.

20

19. Computerprogrammprodukt, das bei seinem Laden und Ausfüh-  
ren auf einem Computer (12) ein Verfahren nach einem der  
Ansprüche 1 bis 15 bewirkt.

) 25